



上海交通大学

SHANGHAI JIAO TONG UNIVERSITY



上海交通大学外国语学院
School of Foreign Languages

从语言到数据

大模型时代的语料库建设 与多语言跨文化评测

聂耳聪 Ercong Nie
上海交通大学外国语学院

2026年10月



报告提纲

背景

语言科学与人工智能的双向赋能

语言数据的变化

新型语料库的内涵、特点与建设需求

训练与评测

模型训练学习需要什么语言数据，模型能力检验需要什么语言数据

多语言跨文化评测 研究案例

BMIKE-53、XCOMPS与多语言数据再利用；跨文化多模态食物评测



研究背景与个人简介



聂耳聪

上海交通大学外国语学院 助理教授



德国慕尼黑大学**计算语言学**博士 (summa cum laude)

研究方向为**自然语言处理**和**语言科学智能**，聚焦**语言科学与人工智能的双向赋能**，尤其关注**多语智能**、**大语言模型可解释性**、**脑—模型语言认知对齐**等领域。

Using language science to better connect, understand, and build AI.

Using AI to advance language science.

以语言科学为理解智能的入口和连接智能的媒介，以人工智能拓展语言科学。

Multilingual AI

LLM Interpretability

Language & Cognition

AI for Humanities

Language Science for AI (LS4AI) & AI for Language Science (AI4LS)



上海交通大学外国语学院
School of Foreign Languages

LS4AI

语言科学帮助理解与改进AI

01

Language Science for AI

We use linguistic knowledge, multilingual evidence, and human language-processing research to explain how language models work, diagnose cross-lingual and cross-cultural failures, and design evaluations that better reflect human language behavior.

- Mechanistic interpretability of multilingual LLMs
- Cross-lingual knowledge, transfer, and language confusion
- Human-inspired evaluation and model design

AI4LS

AI提供语言与认知研究工具

02

AI for Language Science

We use language models and computational methods as research instruments for linguistics, cognition, historical language research, digital humanities, and neighboring human and social sciences.

- Computational models of language and cognition
- Low-resource and historical language technologies
- AI-assisted research in the humanities and social sciences

语言数据连接研究问题、实验方法与模型实证



最新国社科语言学选题中的AI交叉研究动向

AI作为研究对象

AI作为研究工具

模型的语言表现
概念与知识表征
人机语言能力比较

语料处理与资源建设
计算建模与语言分析

以人工智能为研究对象的代表性选题包括：

- 《类型学特色构式的人机语言能力对比与涌现机制研究》
- 《人类语言加工参照下的大语言模型跨语言表征机制研究》
- 《汉语复合逻辑构式的人机理解差异及认知机制研究》
- 《面向人工智能治理的大语言模型元语言能力评测体系构建研究》

这些课题将机器语言能力与人类语言能力进行比较，为语言学研究提供了新的经验对象。机器的语言表征与人类语言知识之间的关系、大语言模型在不同语言任务中的表现及其理论意义、人机语言加工差异所反映的认知问题，已经进入语言学研究的视野。

三、人工智能正在拓展语言学的研究对象与问题空间

从研究内容来看，人工智能进入语言学主要体现为两种路径：一是作为研究工具，参与语言材料处理、语义分析、知识表示和计算建模；二是作为直接研究对象，其语言表征、理解能力和生成机制成为语言学研究的新议题。

与此同时，一些项目以语言学问题为核心，利用人工智能扩展研究方法，例如：

- 《大语言模型辅助的汉语感官词语义地图构建与研究》
- 《基于大语言模型的汉语名词谓语句语体—功能关联研究》
- 《汉语关联副词的理论体系建构及垂直大模型开发》

两种研究路径分别体现了人工智能作为语言研究对象和研究方法的学术价值。前者拓展了语言能力研究的比较维度，后者增强了语言数据处理、知识表示及模型建构的技术条件。

人工智能正在推动语言学的研究工具、研究对象与问题意识共同拓展。人机语言能力比较、语言表征机制、智能语言评价及模型语用能力等议题，为理论语言学、计算语言学与认知科学之间的深入融合提供了新的研究空间。



各类分析也都关注到AI在语言学研究中的双重角色

4.2 研究形态：一半"研究AI的语言"，一半"用AI研究语言"

117项选题呈现出一个漂亮的二分结构：

①本体型（49项，41.9%）

把机器语言本身当作研究对象。如"大模型生成'幻觉文本'的语言哲学检视与治理研究"（黑龙江大学）、"面向人工智能治理的大语言模型元语言能力评测体系构建研究"（南京邮电大学）、"人类语言加工参照下的大语言模型跨语言表征机制研究"（上海交通大学）。

②工具型（58项，49.6%）

把AI当作研究手段。如古文字数据库、方言语料库、多模态临床语料库、机器学习辅助的二语诊断等。

③其他（10项）

数字化教育生态等外围选题。

风头最盛的，当属“**语言+**”的交叉赛道，最耀眼的莫过于语言学与人工智能的联姻。近年人工智能迅猛发展，语言学者紧跟潮流，在题目中嵌入“**生成式人工智能**”字样。此外，高频的关键词还有：**人机协同、语言障碍、康复、失语症、孤独症、数智、数字化、青少年抑郁症、养老、老年语言**，等等。

最为凸显的，恐怕是数智语言研究。立项的课题大多围绕大模型、中文智能评测、人机对话、面向AI的句法资源建设展开。这类题目意味着，现在的语言学研究要跳出单纯地研究人如何表达的传统范式，进入到研究机器如何生成语言，以及机器语言如何反作用于人类的表达习惯。语言学不再局限于人的内部语言，更要研究人与AI的关系。



两条研究主线：多语言 × 可解释性

1

大模型多语言跨文化感知的机制解释与评测
研究大语言模型如何获取、表示、编辑、检索和使用不同语言中的语言知识与事实知识，并解释语言混淆、跨语言不一致和跨文化偏差等现象。

2

人类语言认知加工与大模型机制可解释性的对比研究

将心理语言学、神经语言学和认知科学中的实验范式与模型内部表征分析相结合，比较人类与大模型在形式、意义、知识和推理过程中的异同。

文化
感知

多语言

可解释性

人类语言
加工参照

语言数据需求：可比的跨语言材料与人类语言加工数据



广义语言数据：语言及其加工过程

语言产出

书面文本、口语转写
对话、手语、标注材料

多模态语境

图像、音频、视频
与语言在时间或语义上对齐

人类语言加工

反应时、眼动、脑电、脑成像
实验刺激、条件与被试反应

任务与交互过程

工具调用、环境反馈
智能体行为与人类交互中的语言轨迹数据



这场报告关注的语言数据

面向大模型时代的语言数据

从文本扩展到多模态



怎样帮助模型学习

提供知识、示范、反馈
和任务执行经验



怎样检验模型

设计可比、可控的材料
揭示模型能力

以多语言与跨文化研究为例，讨论语言数据怎样成为研究证据



语料库建设的发展脉络

“大规模”与“精加工”

语料库建设的“初心”是面向语言教学和语言研究，同时面向自然语言处理技术研发。

1960—80年代

1990年代

2010年代

大模型时代

国外 BROWN、LOB

约百万词
平衡采样

BNC、PTB树库

规模扩大（1亿词/3M）
句法等精细标注

网络与多语语料

网页采集
跨语言资源整合

多源、多模态数据

训练、对齐、评测
任务与过程数据

国内 JDEST

百万词规模
第一个科技英语语料库

CCL、BCC

CCL约2700万字
BCC约62亿字

“深度学习时代以来，语料库已演变为支撑人工智能基本能力迭代并迈向通用人工智能的核心要素；没有语料库就没有语言大模型”

趋势：对大规模“生”语料库的需求呈爆炸式上升，以人工标注手段为主的语言“精加工”则从语料库建设的主导地位退居至“不太起眼”的位置。



新型国家语料库的政策背景

《教育强国建设规划纲要（2024—2035年）》

提出建设新型国家语料库

《教育发展“十五五”规划》

加强数字中文建设，推进新型国家语料库建设和应用

语言资源建设成为支撑智能化发展的基础工作

https://www.moe.gov.cn/jyb_xgk/moe_1777/moe_1778/202501/t20250119_1176193.html
https://www.moe.gov.cn/jyb_xgk/moe_1777/moe_1778/202606/t20260629_1442045.html



新型语料库的五个特点

规模极其浩大

数据量级增长（从均衡采样到“全空间高质量”内容收录）
DeepSeek-V3训练语料库14.8万亿词

范围极其广博

网页、百科、图书、论文、代码等多源内容
从新闻、文学作品等**典型**类型到**多源**类型

语言极其多样

多语言，也包括方言、古籍等语言资源
CC100、mC4、OSCAR、NLLB等覆盖百种语言的多语语料库项目

模态极其丰富

文本与图像、音频、视频的对齐和融合
以文本模态为轴的多模态语料库

加工极其专精

指令、问答、偏好、合成数据及相应质量控制
针对大模型训练和评测的新型数据，高质量数据，合成数据，数据治理技术



“大规模”之外，数据覆盖也在变化

传统语料库

围绕研究目的进行采样
重视体裁平衡与语言标注
支持检索、统计和现象分析



面向大模型的数据集合

覆盖更多来源与知识类型
组织多语言、多模态材料
按训练和评测任务加工

覆盖充分、样本可比、来源可追溯，仍然是基本要求



新型语料库的建设难点

高质量供给不足

优质线下资源数字化，结合合成与增强方法

数据体系不完整

整合多源内容，补充多模态与语言多样性

资源难以共享

明确授权方式，按数据属性设计开放与使用机制

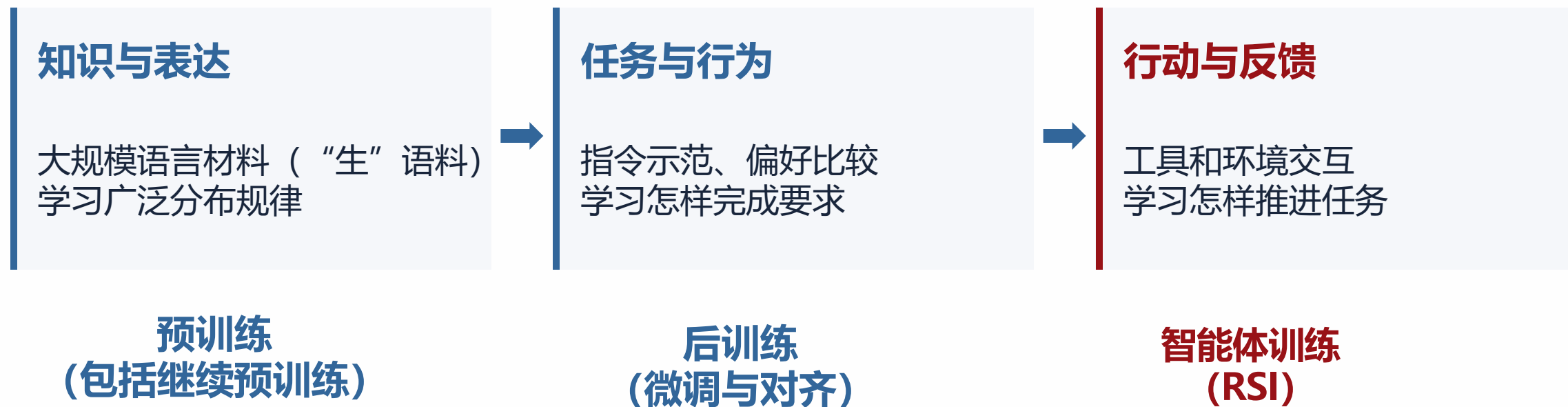
质量标准不统一

贯通采集、清洗、标注、质检与版本管理

从中文语料库建设的角度，数据治理机制和面向大模型时代的“全空间高质量”中文语料库体系亟待完善。



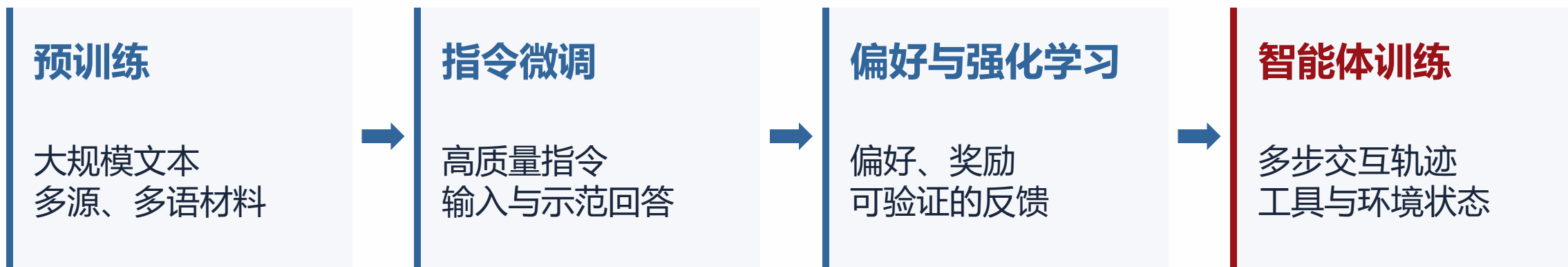
语言数据在模型训练中的作用



训练目标不同，一条训练数据样本的数据形态也不同



训练方式变化，数据需求随之扩展



数据单位

片段 / 文档 (“生” 语料)

指令与回答 (标注数据对)

候选回答与反馈

完整任务过程 (轨迹数据)



预训练与领域适配的数据需求

通用预训练

覆盖语言、知识与表达方式
数据重点：规模、广度、去重与质量过滤

继续预训练 (领域数据)

让模型接触某类专门知识与表达
数据重点：知识密度、领域覆盖与来源质量

监督微调 (SFT)

用任务示范调整回答方式和执行行为
数据重点：指令清楚、答案可靠、任务有代表性

已有基础模型让更多研究者可以聚焦数据设计与能力适配



后训练：从指令跟随到反馈信号

思维链 (CoT) 学习

给出任务与高质量回答
以及**推理过程**
学习可借鉴的完成方式

偏好学习

比较多个候选回答
学习哪些表现更合适

强化学习

生成、验证、获得奖励
根据结果调整策略

高质量任务种子



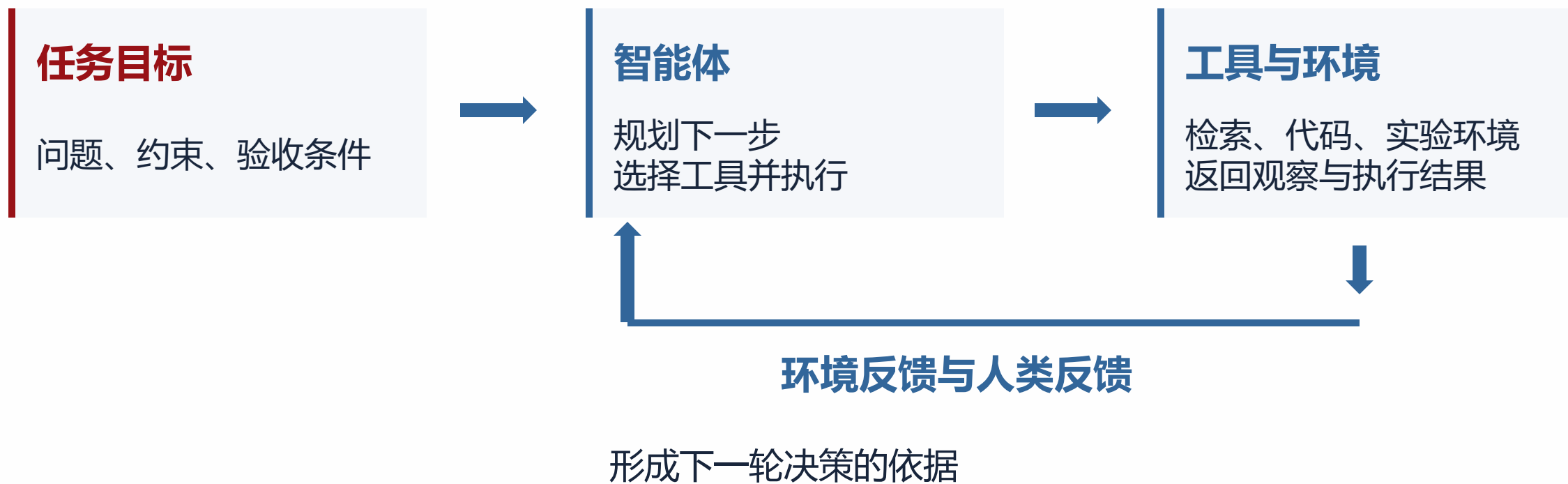
大量候选与交互



筛选与有效反馈



智能体的学习过程形成一个交互闭环



训练数据开始记录任务如何推进，而不只保存最后的答案



一条交互轨迹包含哪些数据

- | | | |
|----|----|---------------|
| 01 | 任务 | 修复一个程序错误 |
| 02 | 动作 | 读取代码，调用测试工具 |
| 03 | 观察 | 测试失败，返回报错信息 |
| 04 | 修正 | 修改相关代码并再次测试 |
| 05 | 反馈 | 测试通过，人类确认符合需求 |

还需保留时间顺序、工具参数、环境版本和成功 / 失败状态



高价值轨迹为什么难采集

复杂任务需要长过程

数学证明、软件开发、复杂的资料分析与论证

过程质量需要验证

最终成功，未必意味着每一步都值得学习

反馈依赖环境与专家

工具状态、验收标准、人类判断影响数据价值

失败也有信息量

错误出现在哪里，怎样发现，怎样恢复

一条高质量的轨迹数据，价值上千美元。



语言数据在模型评测和机制解释中的作用

训练数据帮助模型获得能力

评测数据让能力变得可观察、可比较

会在哪些条件下成功？

换一种语言、语境或模态，还能保持表现吗？

发生错误时，问题出在哪里？



能力边界需要沿多个维度观察

能力维度

语言学能力、知识更新、概念理解、跨文化理解

变化条件

语言、书写系统、文化语境、模态组合、任务难度

观测方式

回答正确性、概率偏好、内部表征、完整执行轨迹

对照依据

可计算指标、最小对、人工判断、可验证结果



最小对：用一个差异检验一个问题

概念属性理解

可接受句

打蛋器可以把空气打入混合物。

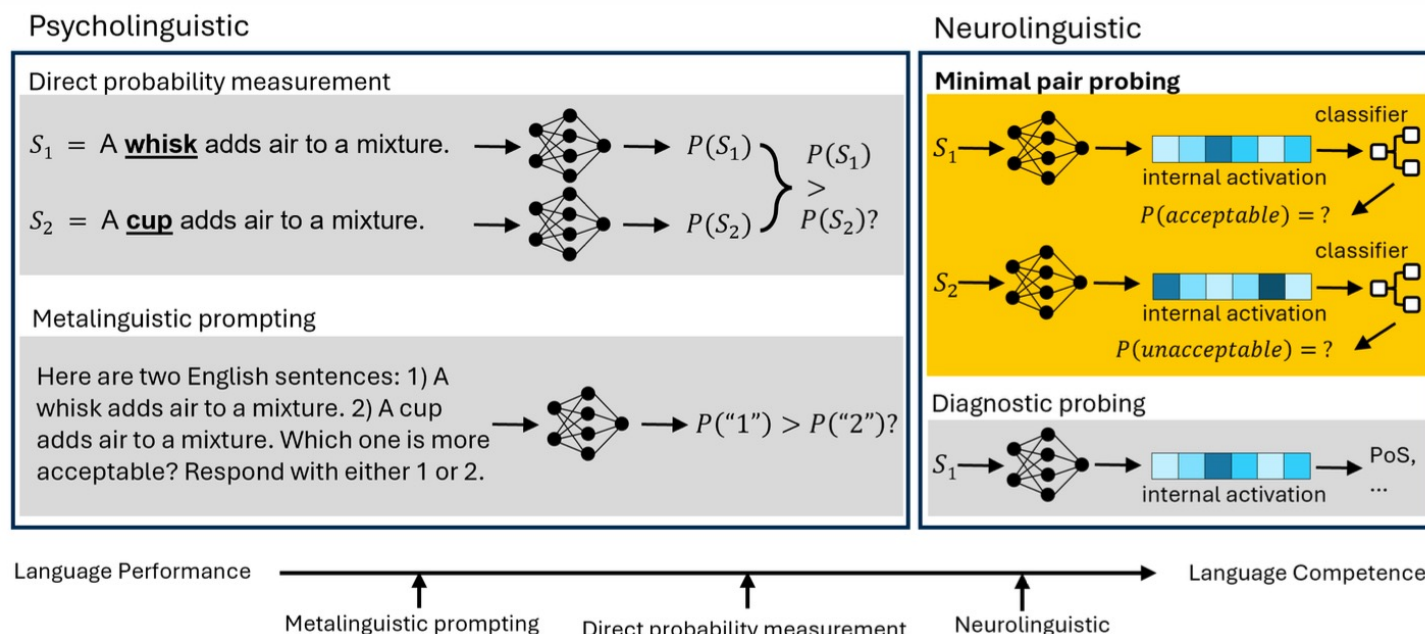
对照句

杯子可以把空气打入混合物。

保留句式与属性，只替换被考察的概念

控制材料差异，便于更清楚地观测和解释模型表现差异

最小对设计数据评价和解释模型



行为

回答是否正确
不同语言是否一致

机制

信息在哪一层可解码
哪些神经元参与加工

同一组语言材料，可以连接输出行为的评价与内部表征的观测



国内外评测任务大会

中国计算语言学大会 (CCL) 评测任务

一、自然语言处理基础任务

任务一：第二届中文叙实性推理评测

任务二：非字面义翻译和理解评测

任务三：跨主流语言与低资源语言对齐的大模型金融评测

任务四：低资源缅甸语固定模版语句推理评测

二、自然语言处理应用

任务五：杂粮育种信息抽取评测

任务六：中文电子病历疾病与手术ICD自动编码评测

任务七：跨语言文学文本情感分析一致性评测

任务八：大语言模型生成中文医疗内容的循证事实核查

任务九：第二届古诗词赏析评测

三、生成式AI与大模型核心能力

任务十：基于情景的常识推理评测

任务十一：面向自动驾驶的自动化危害分析与风险评估评测任务

任务十二：面向听障群体的信息无障碍结构化字幕生成评测

任务十三：图像文本翻译质量评测

任务十四：中文会话含义与隐喻能力评测

ACL SemEval评测任务大会

Reasoning and Inference

- **Task 1: RETECO: A SemEval-2027 Shared Task on Reasoning-Oriented Retrieval** ([contact organizers], [join task mailing list])
Abdelrahman Abdallah, Mohammed Ali, Muhammad Abdul-Mageed, Kevin Duh and Adam Jatowt.
- **Task 2: Directional-Consistent Fine-Grained Natural Language Inference (DiCo-NLI)** ([contact organizers])
Jon Felix Apaolaza Larraya, Aitor Soroa, Rodrigo Agerri and Inigo Lopez-Gazpio

Lexical and Compositional Semantics

- **Task 3: Semantic Change Across Multiple Time Periods** ([contact organizers], [Codabench])
Nina Tahmasebi, Pierluigi Cassotti, Felix Morger, Lucia Siciliani, Eduardo Calò, Pablo Mosteiro, Stefano De Pascale and Mariia Fedorova
- **Task 4: Par-PARSEME: A Multilingual Shared Task on Paraphrasing of Idioms** ([contact organizers])
Carlos Ramisch, Manon Scholivet, Agata Savary, Takuya Nakamura and Eric Bilinski
- **Task 5: ComPartMENT - Noun Compound and Particle Verb Compositionality Prediction in Context** ([contact organizers][Codabench])
Filip Miletić, Chris W. Jenkins and Sabine Schulte im Walde

Cultural and Cross-Lingual Understanding

- **Task 6: Multilingual and Multimodal Culturally Grounded Spoken and Visual Question Answering** ([contact organizers], [slack channel])
Firoj Alam, Shammur Absar Chowdhury, Md Arid Hasan, Nadir Durrani, Fahim Dalvi, Ebru Arisoy and Usman Naseem
- **Task 7: CLaS: Cross-Lingual Alignment and Steering of Large Language Models** ([contact organizers])
Daniil Gurgurov, Tanja Baeumel, Dana Arad, Anastasiia Sedova, Masha Fedzechkina, Josef van Genabith, Patrick Schramowski, Marius Mosbach and Simon Ostermann



多语言评测研究案例

BMIKE-53

跨语言知识编辑

一种语言中提供的新知识，
能否在其他语言中使用？

XCOMPS

跨语言概念理解

换一种语言表达，
概念与属性的关系是否稳定？

BMIKE-53: 跨语言知识编辑

53种语言

zsRE、CounterFact、WikiFactDiff

在一种语言中给出知识更新
在其他语言中检验更新能否迁移
同时检查无关知识是否保留

zsRE

<i>Edited (en)</i> <i>Knowledge</i>	What war did Carlos W. Colby fight in? Korean War
<i>(zh)</i> <i>Test Query</i>	卡洛斯·W·科尔比参与了哪两个国家之间的冲突? <i>Which conflict between two countries did Carlos W. Colby participate in?</i>

CounterFact

<i>Edited (en)</i> <i>Knowledge</i>	In which continent is Shinnan Glacier located? Europe
<i>(zh)</i> <i>Test Query</i>	新南冰川所在大陆的最高峰是哪座山? <i>Which mountain is the highest peak on the continent where the Shinnan Glacier is located?</i>

WikiFactDiff

<i>Edited (en)</i> <i>Knowledge</i>	For which team does Masaki Yamamoto play? Team Ukyo
<i>(zh)</i> <i>Test Query</i>	山本雅树效力的团队的老板是谁? <i>Who is the owner of the team for which Masaki Yamamoto plays?</i>

该研究的是上下文知识编辑 (IKE) , 无需更新模型参数



BMIKE-53的数据构建



多语言扩展同时要保留实体、事实关系和任务标签的一致性

数据再利用的关键是“保留任务”，而不只是翻译句子



BMIKE-53: 一个编辑对应四类检验

上下文中的知识更新

用源语言给出待更新的事实



可靠性

目标问题能否采用新知识

泛化性

换一种问法是否仍然有效

局部性

无关知识能否保持不变

可迁移性

相关推理问题能否用上新知识

跨语言编辑的成功，需要同时考虑更新效果与知识保留

BMIKE-53: 一条检测数据示意图

An example of cross-lingual IKE

Edited knowledge (*source language - en*):

When was Darrell Spencer born? 1944

Locality query (*target language - de*):

Bei welchem historischen Ereignis wurde Darrell Spencer geboren?

In which historical event was Darrell Spencer born?

Prompt construction

Zero-shot cross-lingual IKE

New fact: "When was Darrell Spencer born? 1944"

Questions: "Bei welchem historischen Ereignis wurde Darrell Spencer geboren?" Answer:

One-shot cross-lingual IKE

New fact: "What did Earl Hooker die of? tuberculosis"

Question: "Was war die Todesursache von Earl Hooker?" Answer: "Tuberkulose"

New fact: "When was Darrell Spencer born? 1944"

Questions: "Bei welchem historischen Ereignis wurde Darrell Spencer geboren?" Answer:

Few-shot cross-lingual mixed IKE

Demo 1 Demo 2 Demo 3 Demo 4 Demo 5 Demo 6 Demo 7 Demo 8

New fact: "When was Darrell Spencer born? 1944"

Questions: "Bei welchem historischen Ereignis wurde Darrell Spencer geboren?" Answer:

Few-shot cross-lingual metric-specific IKE

Demo 1 Demo 2 Demo 3 Demo 4 Demo 5 Demo 6 Demo 7 Demo 8

New fact: "When was Darrell Spencer born? 1944"

Questions: "Bei welchem historischen Ereignis wurde Darrell Spencer geboren?" Answer:

Types of In-context Learning Demonstrations

Reliability Demo

New fact: "In what war did Jesse Orin Creech fight in? World War I"

Question: "In welchem Krieg hat Jesse Orin Creech gekämpft?" Answer: "Erster Weltkrieg"

Question: "In what war did Jesse Orin Creech fight in?" Answer: "World War I"

Generality Demo

New fact: "What did Earl Hooker die of? tuberculosis"

Question: "Was war die Todesursache von Earl Hooker?" Answer: "Tuberkulose"

Question: "What was the cause of Earl Hooker's death?" Answer: "tuberculosis"

Locality Demo

New fact: "What position does Malcolm Partridge play? Forward"

Question: "Wer war der Quarterback der Wikinger 1993?" Answer: "Jim McMahon"

Question: "Who was the Vikings' quarterback in 1993?" Answer: "Jim McMahon"

Portability Demo

New fact: "Who is the architect that designed Toodyay Fire Station? Wyndham Lewis"

Question: "Mit welcher Kunstbewegung ist der Architekt von Toodyay Fire Station in Verbindung gebracht?" Answer: "Wirbel"

Question: "What art movement is the architect of Toodyay Fire Station associated with?" Answer: "vorticism"

Zweiter Weltkrieg
World War II



BMIKE-53的主要观察

模型规模

较大模型通常更好地处理跨语言迁移与知识保留

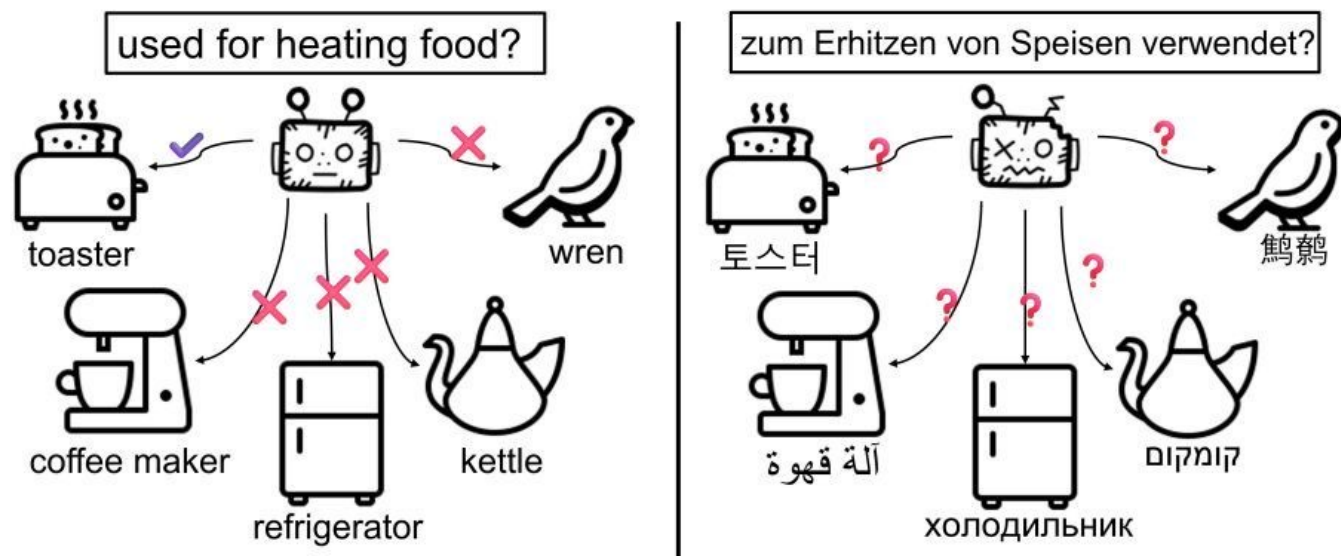
上下文示范

示范内容与评测指标的匹配，会显著影响编辑效果

语言与书写系统

非拉丁文字语言表现更弱，语言混淆是相关问题之一

XCOMPS: 跨语言概念理解



17种语言

521个概念

同一概念与属性
在不同语言中对齐

“概念” + “属性”

例如：

“烤面包机用于加热食物”能否在不同语言中被正确理解？

概念：“烤面包机”；属性：“用于加热食物”



XCOMPS的四类负例

Type	Language	Acceptable Sentence	Unacceptable Sentence
Taxonomic	Spanish	<i>Tostadora</i> se utiliza para calentar alimentos. (A toaster is used for heating food.)	<i>Cafetera</i> se utiliza para calentar alimentos. (A coffee maker is used for heating food.)
Overlap	Vietnamese	<i>Máy nướng bánh mì</i> được sử dụng để hâm nóng thực phẩm. (A toaster is used for heating food.)	<i>Tủ lạnh</i> được sử dụng để hâm nóng thực phẩm. (A refrigerator is used for heating food.)
Co-occurrence	Hungarian	<i>Kenyérpirító</i> ételek melegítésére használják. (A toaster is used for heating food.)	<i>Vízforraló</i> ételek melegítésére használják. (A kettle is used for heating food.)
Random	Dutch	<i>Broodrooster</i> wordt gebruikt om voedsel te verwarmen. (A toaster is used for heating food.)	<i>Winterkoning</i> wordt gebruikt om voedsel te verwarmen. (A wren is used for heating food.)

Table 1: XCOMPS examples, illustrating each linguistic variant pairs an acceptable sentence (positively matched property) with an unacceptable counterpart (negatively matched property).

分类相近
Taxonomic

属性重叠
Overlap

经常共现
Co-occurrence

随机概念
Random

负例越接近，越能检验细微的语义区分



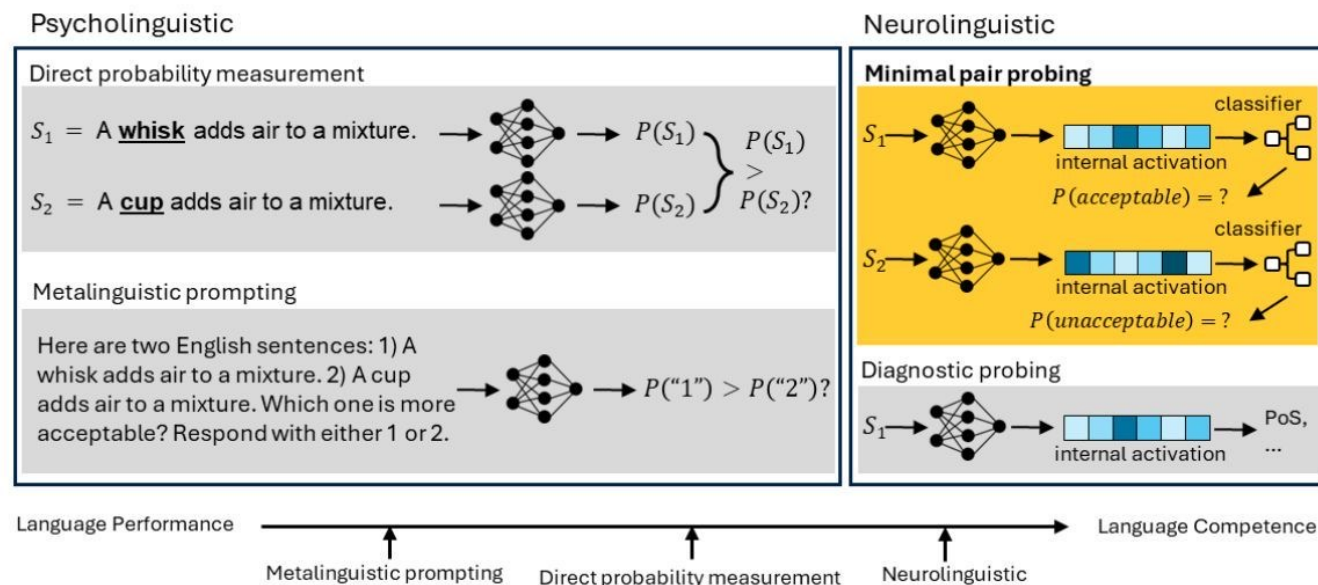
XCOMPS的人机协同构建流程



例如 bat 的含义需要先确定，再选择“蝙蝠”或“球拍”等目标语言表达

多语种子帮助消歧，人工核查保证概念对应

同一批材料，三种评测视角



显式判断 (Meta)

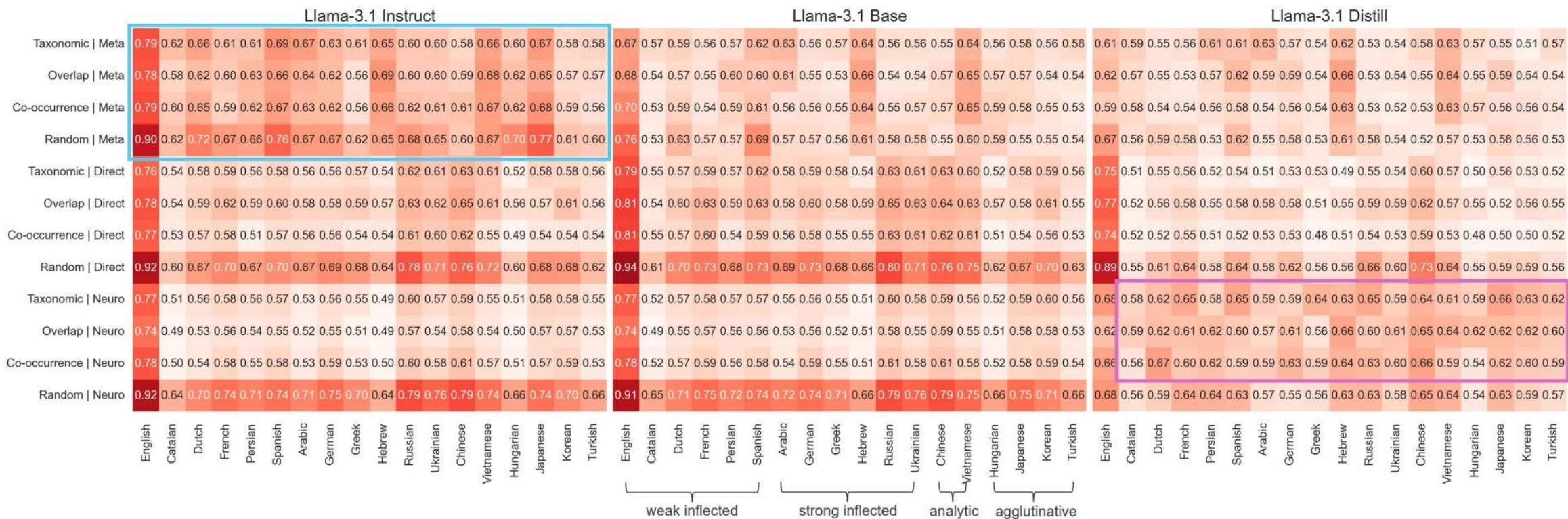
概率比较 (Direct)

内部探测 (Neuro)

内部可解码信息提供补充证据，不能直接等同于人类语言能力

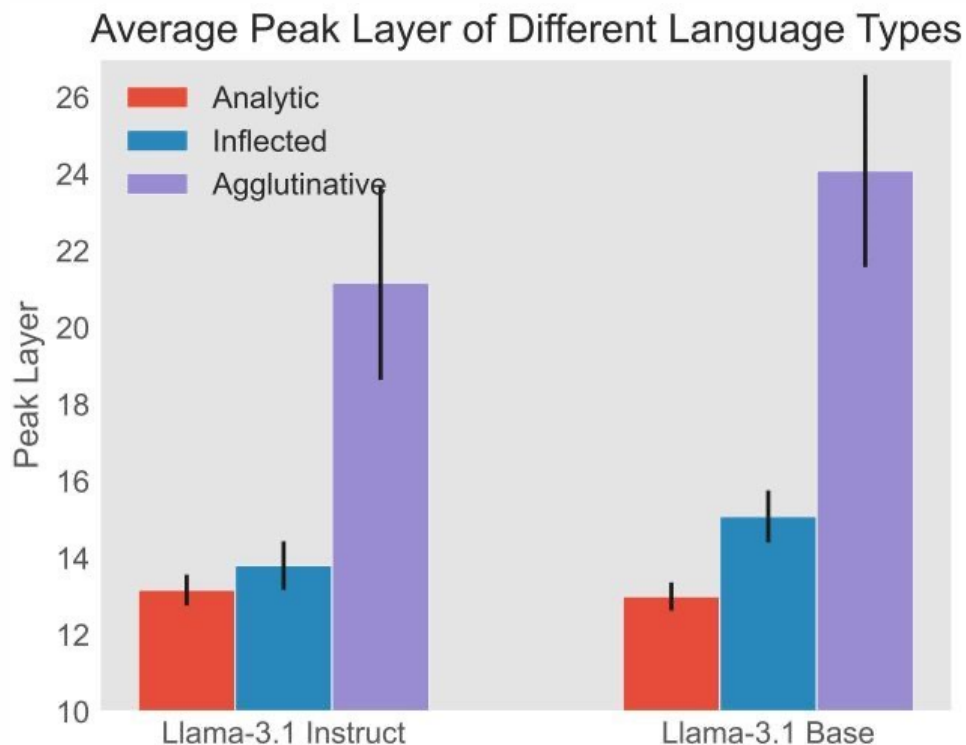


XCOMPS: 表现随语言和负例而变化



随机负例相对容易，语义接近的负例更难；平均分会掩盖语言差异

语言类型与模型内部表征



不同语言类型

探测表现及达到峰值的层位置
存在系统差异

解释时还需考虑

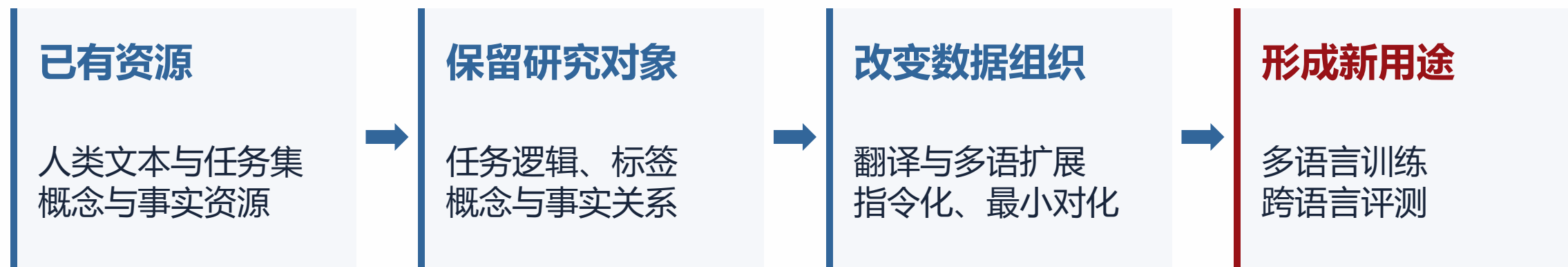
训练数据、分词方式与语言覆盖
等因素的共同影响

语言数据让类型学问题转化为可比较的模型实验



Data Repurposing

数据再利用：让已有资源支持新问题



BMIKE-53与XCOMPS都体现了“资源继承 + 任务重组”的思路



多语言扩展后的质量检查

意义是否保留

歧义是否消除，概念和事实有没有变

任务是否保留

翻译之后，标签是否仍成立，难度是否发生变化

表达是否自然

是否出现翻译腔、固定模板或生成模型的语言习惯

文化是否适用

目标文化中是否有同样的对象、经验和默认前提

覆盖更多语言之后，还需要验证数据是否仍然可比



多语言能力与跨文化理解

语言扩展

同一问题用不同语言表达
比较知识与能力能否迁移

文化扩展

纳入不同地方的知识与实践
检验模型怎样理解具体语境

翻译能扩展语言覆盖，文化覆盖还需要来自当地的材料

食物文化：图像里还有哪些信息



粽子



Tamales



Gumbo

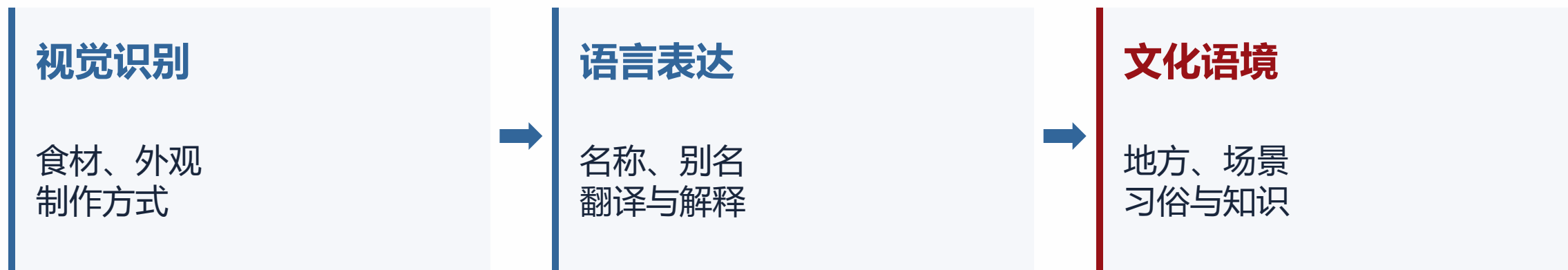


月饼

叫什么？来自哪里？什么时候吃？有哪些地方做法与文化含义？



跨文化多模态评测：以食物为例



识别出一道食物，和理解它的文化语境，是不同层次的任务

WorldCuisines的多语言图文任务

30种语言及方言

9个语系

100万以上图文数据点

菜名识别、地点判断
多选与开放回答
有语境与无语境条件

Task 1a: Dish name prediction (with corresponding translation example)



Type: **Multiple-Choice** English

What is this culinary dish called?

1) Strudel 4) New York roll
2) Mille-feuille 5) Milk roll
3) Milk-cream strudel

Print only the answer with a single answer id (1,2,3,4,5).

Type: **Open-Ended** English

What is this culinary dish called?

Print only the answer.

Type: **Multiple-Choice** Japanese (formal)

この料理は何と言いますか?

1) シュトゥルーデル 4) New York roll
2) ミルフィーユ 5) Milk roll
3) ミルヒラーム・シュトゥルーデル

Print only the answer with a single answer id (1,2,3,4,5).

Type: **Open-Ended** Japanese (casual)

この料理、何？

Print only the answer.

Task 1b: Contextualized dish name prediction



Type: **Multiple-Choice** English

What is the common name for this dish in Brazil?

1) Paçanga böreği 4) Sopaipilla
2) Coxinha 5) Empadão
3) Qottab

Print only the answer with a single answer id (1,2,3,4,5).

Task 1c: Dish name prediction with adversarial context



Type: **Multiple-Choice** English

Yesterday I had a nice lunch at a Korean restaurant. I am about to have this dish now. What is this dish called?

1) Scotch woodcock 4) Gyeran-ppang
2) Matzah brei 5) Oeuf mayonnaise
3) Eggs Benedict

Print only the answer with a single answer id (1,2,3,4,5).

Task 2: Location prediction



Type: **Multiple-Choice** English

Which country made this dish popular?

1) Poland 4) Greece
2) France 5) Germany
3) Finland

Print only the answer with a single answer id (1,2,3,4,5).

Type: **Multiple-Choice** French

Quel est le pays qui a rendu ce plat populaire?

1) Pologne 4) Grèce
2) France 5) Allemagne
3) Finlande

Print only the answer with a single answer id (1,2,3,4,5).

同一张图片，不同的语境条件

无语境

图片 + 问题
观察视觉与已有知识



有帮助的语境

图片 + 有效线索
检验语境利用能力



误导性语境

图片 + 干扰线索
检验抗干扰能力

对比条件

保持图片与目标问题一致，改变附加的文字线索

多模态评测需要同时检查图像信息和文本语境怎样影响判断

同一张图片，不同的语境条件（例子）

1. No context



Image only



AI reasoning

Not enough information.
Could be many dishes.



AI answer

Maybe fried rice or noodles?
(Uncertain)

2. Correct context



Image + helpful context



Region:
South Asia



Cuisine:
Indian



Clues:
Spices, herbs, basmati rice



AI reasoning

Rice + spices + Indian
cuisine → likely a
spiced rice dish.



AI answer

Vegetable Biryani
(Likely correct)

3. Misleading context



Image + misleading context



Region:
East Asia



Cuisine:
Chinese



Clues:
Soy sauce, wok-cooked



AI reasoning

Wok-cooked + soy sauce
→ likely a **Chinese**
fried rice dish.



AI answer

Chicken Fried Rice
(Likely incorrect)



The same image can lead to different answers depending on context.



从语言到数据：总结

面向大模型时代的 语言数据

文本、多模态、人类语言加工与任务交互过程

模型训练的数据需求

规模之外，更关注质量和领域，以及随着不同任务需求的新数据类型
最近智能体训练开始需要高质量轨迹数据

语言数据支持模型 评测

控制变量、对齐材料、分析语境，解释能力边界

多语言跨文化评测

用可比较的数据检验模型，也让数据建设回应真实差异

谢谢！欢迎交流



参考文献：背景与数据建设

孙茂松、孔存良. 2026. 中文语料库建设呼唤新格局. 光明日报, 7月5日. 清华大学转载:
<https://www.tsinghua.edu.cn/info/1662/127334.htm>

全国哲学社会科学工作办公室. 2026. 2026年国家社会科学基金年度项目立项名单公示及附件1（重点项目）. 10月8日. <https://www.nopss.gov.cn/n1/2026/1008/c431027-40810871.html>

中共中央、国务院. 2025. 教育强国建设规划纲要（2024—2035年）. 1月19日.
https://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202501/t20250119_1176193.html

国务院. 2026. 教育发展“十五五”规划. 教育部网站, 6月29日.
https://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202606/t20260629_1442045.html



参考文献：训练与智能体

DeepSeek-AI, Guo, D., Yang, D., et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948. <https://arxiv.org/abs/2501.12948>

Luo, X., Zhang, Y., He, Z., Wang, Z., Zhao, S., Li, D., Qiu, L. K., & Yang, Y. 2025. Agent Lightning: Train ANY AI Agents with Reinforcement Learning. arXiv:2508.03680. <https://arxiv.org/abs/2508.03680>



参考文献：多语言与模型分析

Nie, E., Shao, B., Wang, M., Ding, Z., Schmid, H., & Schütze, H. 2025a. BMIKE-53: Investigating Cross-Lingual Knowledge Editing with In-Context Learning. ACL, 16357–16374.
<https://doi.org/10.18653/v1/2025.acl-long.798>

He, L., et al. 2025a. XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs. SIGTYP.
<https://aclanthology.org/2025.sigtyp-1.9/>

He, L., Nie, E., Schmid, H., Schütze, H., Mesgarani, N., & Brennan, J. 2025b. Large Language Models as Neurolinguistic Subjects: Discrepancy between Performance and Competence. Findings of ACL, 19284–19302. <https://doi.org/10.18653/v1/2025.findings-acl.986>

Nie, E., et al. 2025b. Mechanistic Understanding and Mitigation of Language Confusion in English-Centric Large Language Models. Findings of EMNLP. <https://aclanthology.org/2025.findings-emnlp.37/>



参考文献：概念与跨文化评测

Misra, K., Rayz, J., & Ettinger, A. 2023. COMPS: Conceptual Minimal Pair Sentences for Testing Robust Property Knowledge and Its Inheritance in Pre-trained Language Models. EACL, 2928–2949.

Winata, G. I., et al. 2025. WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines. NAACL. <https://doi.org/10.18653/v1/2025.naacl-long.167>

Li, W., et al. 2024. FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture. EMNLP. <https://doi.org/10.18653/v1/2024.emnlp-main.1063>

He, L., Nie, E., et al. 2026. Investigating Multilingual Conceptual Understanding of Large Language Models with Minimal Pair Probing